STAT2_HW1

Xingyu Chen

February 07, 2022

Total Score: 42/48

1 Homework #1

See Canvas for HW #1 assignment due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTex/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy. Please do not turn in messy work. Computational problems should be completed in this notebook (using the R kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

1.1 A. Theoretical Problems

1.1.1 Problem A.1

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known, and we are interested in an estimator for $\theta = \mu^2$. (Note that we will use the following calculuations to make comparisons to the parametric boostrap method explored below).

(a) Find the maximum likelihood estimator (MLE) for θ, denoted $\hat{\theta}$. Q1: 2/2

Answer:

The pdf of normal distribution is:

$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The MLE for μ^2 will be:

$$L(\mu^2 \mid \theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \prod_i^n e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Take the log for both side:

$$\ln L(\mu^2 \mid \theta) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{\sum(x-\mu)^2}{2\sigma^2}$$

The take the derivative for both side:

$$\frac{\partial \ln L(\mu^2 \mid \theta)}{\partial \mu^2} = \frac{1}{\sigma^2}\sum_i^n (x_i - \mu) \equiv 0$$

Thus,

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_i^n x_i$$

$$\hat{\theta} = \hat{\mu^2} = (\frac{1}{n}\sum_i^n x_i)^2$$

(b) Compute the bias of $\hat{\theta}$, denoted Bias($\hat{\theta}$). Recall that Bias($\hat{\theta}$) $= E(\hat{\theta}) - \theta$. Q2: 2/2

Answer:

2 5

we have:

$$\bar{x} = \sum_i^n \frac{x_i}{n}$$

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\bar{x}^2) - \mu^2 = Var[\bar{x}] + E[\bar{x}]^2 - \mu^2$$

$$= \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

Thus,

$$\text{Bias}(\hat{\theta}) = \frac{\sigma^2}{n}$$

(c) Compute the variance of $\hat{\theta}$, denoted Var($\hat{\theta}$) (HINT: You might use a moment generating function at some point in your answer.) Q3: 2/2

Answer:

2 5

$$\text{Var}(\hat{\theta}) = \text{Var}(\mu^2) = \text{Var}(\bar{x}^2)$$

The MGF of $\bar{x}$ is $e^{\mu t + \frac{t^2\sigma^2}{2n}}$.

Thus,

$$m'(t) = (\mu + \frac{t\sigma^2}{n}) * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m''(t) = (\mu + \frac{t\sigma^2}{n})^2 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{\sigma^2}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m'''(t) = (\mu + \frac{t\sigma^2}{n})^3 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{3\sigma^2(\mu + \frac{t\sigma^2}{n})}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m''''(t) = (\mu + \frac{t\sigma^2}{n})^4 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{6\sigma^2(\mu + \frac{t\sigma^2}{n})^2}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{3\sigma^4}{n^2} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$\text{Var}(\hat{\theta}) = m''''(0) - [m''(0)]^2 = \mu^4 + \frac{6\sigma^2\mu^2}{n} + \frac{3\sigma^4}{n^2} - (\mu^2 + \frac{\sigma^2}{n})^2 = \frac{4\sigma^2\mu^2}{n} + \frac{2\sigma^4}{n^2}$$

(d) Write down the boostrap estimators of Bias($\hat{\theta}$) and Var($\hat{\theta}$). Q4: 1/2

Answer:

From (b), (c) and we know that

$$\hat{x} = \frac{1}{n}\sum_i^n x_i$$

and$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ Thus, the boostrap estimators will be:

$$\text{Bias}(\hat{\theta})_{boot} = \sigma^2/n = \frac{\sum(x_i - \bar{x})^2}{n(n-1)}$$

.

$$\text{Var}(\hat{\theta})_{boot} = \frac{4\sigma^2\mu^2}{n} + \frac{2\sigma^4}{n^2}$$

1.1.2 Problem A.2

Provide a brief explanation of the pros and cons of using the bootstrap for calculating confidence intervals. Q5: 2/2

Answer:

Pros: It is a straightforward way to derive estimates of confidence intervals for complex estimators of the distribution. It is also an appropriate way to control and check the stability of the results. It is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality.

Cons: Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency. The result may depend on the representative sample.

1.2 B. Computational Problems

1.2.1 Problem B.1

Suppose that $X_1, \ldots, X_8 \overset{iid}{\sim} \Gamma(\alpha, \beta)$. Let's use the bootstrap to compute a 90% confidence interval for the population standard deviation: $sd(X) = \sqrt{\alpha/\beta^2} = \theta$.

Note:

The convention in this course will be to interpret $\Gamma(\alpha, \beta)$ as the "shape/rate" parameterization: shape $= \alpha$, rate $= \beta$. But R uses the "shape/scale" parameterization: shape $= \alpha$, scale $= \theta = 1/\beta$.

To be sure that you are properly simulating from the right gamma distribution, see the help file for rgamma() (run: ? rgamma). Also, see here for more information on the gamma distribution.

(a) State why a χ^2 confidence interval is not valid in this context. You should reply on knowledge from your prereq class! Q6: 1/2

Answer:

χ^2 confidence interval is used when the data are normal, not for gamma.

2 +1

(b) Generate a sample of size $n = 8$ from $\Gamma(\alpha = 3, \beta = 4)$ and calculate the true population standard deviation (in this example, we are generating data so that we can see how well our estimation procedure will do). Q7: 2/2

Answer:

2 +1

```
shape = 3
scale = 1/4
n = 8
z_90 = 1.645
std = sqrt(shape*scale^2)
print(std)
```

```
## [1] 0.4330127
```

(c) Generate $B = 200$ bootstrap samples from the above sample. Print the dimension, and articulate what each row/column represents. Avoid loops! (HINT: use the replicate() function.) Q8: 2/2

Answer:

2 +1

```
library(boot)
dat <- rgamma(n=n, shape = shape, scale = scale)
N <- 200
set.seed(1234)
result = replicate(N, sample(dat, replace = T), simplify = 'matrix')
df <- data.frame(t(result))
print(dim(df))
```

```
## [1] 200   8
```

Each row represent one round, each column represent one sample random choose from the data with replacement.

(d) Calculate and print the sample standard deviation, s. Then, calculate s for each bootstrap sample. Denote this as s_i^*, for $i = 1, \ldots, B$. Avoid loops! (HINT: use the apply() function.) Display a histogram of the distribution of $s_i^*, i = 1, \ldots B$ Q9: 2/2

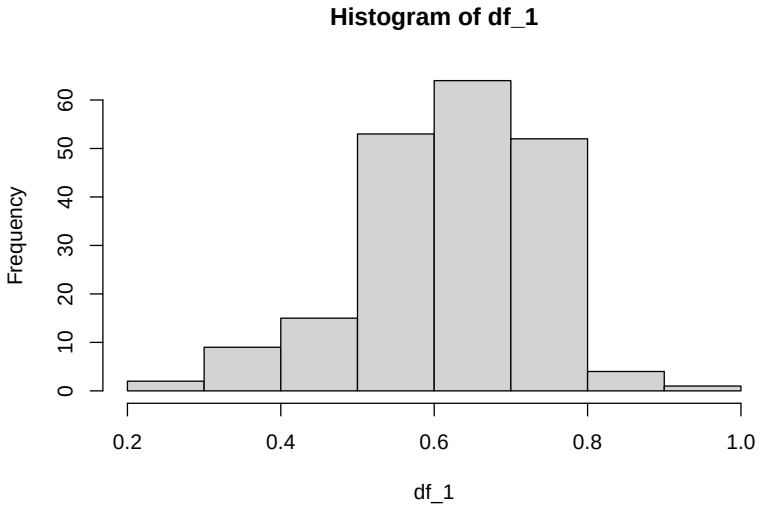
Answer:

2 +1

```
sd(dat)
```

```
## [1] 0.6886518
```

```
df_1 <- apply(df, 1, sd)
hist(df_1)
```



(e) Use the quantile 0 function to find the 5 th and 95 th percentile of the distribution of s_i^*. Use these values to calculate the 90% boostrap pivot confidence interval and bootstrap percentile confidence interval for θ. Q10: 1/2

Answer:

1 +1

```
x_5 = quantile(df_1, 0.05)
x_95 = quantile(df_1, 0.95)
cat("(",x_5,",",x_95,")")
```

```
## ( 0.39836 , 0.7789232 )
```

(f) Interpret this confidence interval. Q11: 2/2

Answer:

2 +1

The interval that I can be 90% certain contains the population standard deviation.

1.2.2 Problem B.2

Thus far, we've been looking at the nonparametric bootstrap. In this problem, we look at the parametric bootstrap as a way of estimating the bias and variance of an estimator $\hat{\theta} = \bar{X}^2$ of $\theta = \mu^2$ (in problem A.1 you calculated these values exactly).

(a) Generate $X_1, \ldots, X_{20} \overset{iid}{\sim} N(\mu = 2, \sigma^2 = 1)$, and then forget that you know μ and σ^2. Find the sample mean and sample variance. Q12: 2/2

Answer:

```
dat_b2 <- rnorm(20, mean = 2, sd =1)
mean(dat_b2)
```

```
## [1] 1.980578
```

```
sd(dat_b2)
```

```
## [1] 0.9583407
```

(b) Define $\widehat{N}$ to be the distribution of the variable X_i in the population with the sample estimates plugged in for the unknown population parameters. Write down $\widehat{N}$ based on the data generated in (a). Q13: 2/2

Answer:

$\widehat{N} = N(\mu = 1.980578, \sigma^2 = 0.9583407)$

(c) Draw $B = 500$ parametric bootstrap samples from $\widehat{N}$, and for each bootstrap sample $(X_{1,j}, \ldots, X_{20,j})$, compute

$$\hat{\theta}_j^* = \left(\frac{1}{20} \sum_{i=1}^{20} X_{i,j}^* \right)^2$$

where $j = 1, \ldots, B$ Q14: 2/2

Answer:

```
mu_b2 <- 1.980578
var_b2 <- 0.9583407
myData <- rnorm(20,mu_b2,var_b2)
set.seed(200) # Setting the seed for replication purposes
sample.size <- 20 # Sample size
n.samples <- 500 # Number of bootstrap samples
bootstrap.results <- c() # Creating an empty vector to hold the results
for (i in 1:n.samples)
{
    obs <- sample(1:sample.size, replace=TRUE)
    bootstrap.results[i] <- mean(myData[obs]) * mean(myData[obs]) # Mean of the bootstr
}
head(bootstrap.results)
```

```
## [1] 4.471069 4.382608 3.655353 5.562506 4.177846 5.534534
```

(d) Compute an estimate of the bias:

$$\widehat{B}(\hat{\theta}) \approx \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^* - \bar{x}^2$$

Compare this to the exact bias using the formula in problem A.1. Q15: 2/2

Answer:

```
print(mean(bootstrap.results) - mu_b2^2)
```

```
## [1] 0.7367678
```

```
print(var_b2^2 / 20)
```

```
## [1] 0.04592084
```

(e) Compute an estimate of the variance:

$$\widehat{\mathrm{Var}}(\hat{\theta}) \approx \frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{\theta}_j^* - \bar{\theta}\right)^2,$$

where

$$\bar{\theta} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$$

Compare this to the exact variance using the formula in problem A.1. Q16: 2/2

Answer:

2 +1

```
print(sum((bootstrap.results - mean(bootstrap.results)) * (bootstrap.results - mean(boot
```

```
## [1] 0.712469
```

```
print((4*(mu_b2^2)*(var_b2^2)/500) + 2*(var_b2^4)/(500^2))
```

```
## [1] 0.02882806
```

(f) True or False: For a fixed sample size $n = 20$, as B increases, $\widehat{\mathrm{Bias}}(\hat{\theta})$ will approach $\mathrm{Bias}(\hat{\theta})$. That is, for a fixed n, the bootstrap estimate of the bias will approach the true bias as the number of bootstrap samples, B increases. You might consider running a simulation to decide! Q17: 2/2

Answer:

False.

2 +1

```
mu_b2 <- 1.980578
var_b2 <- 0.9583407
myData <- rnorm(20,mu_b2,var_b2)
set.seed(200) # Setting the seed for replication purposes
sample.size <- 20 # Sample size
n.samples <- 50000 # Number of bootstrap samples
bootstrap.results <- c() # Creating an empty vector to hold the results
for (i in 1:n.samples)
{
    obs <- sample(1:sample.size, replace=TRUE)
    bootstrap.results[i] <- mean(myData[obs]) * mean(myData[obs]) # Mean of the bootstr
}
print(mean(bootstrap.results) - mu_b2^2)
```

```
## [1] 1.365394
```

```
print(var_b2^2 / 20)
```

```
## [1] 0.04592084
```

```
print(sum((bootstrap.results - mean(bootstrap.results)) * (bootstrap.results - mean(boot
```

```
## [1] 58.65064
```

```
print((4*(mu_b2^2)*(var_b2^2)/500) + 2*(var_b2^4)/(500^2))
```

```
## [1] 0.02882806
```

1.2.3 Problem B.3

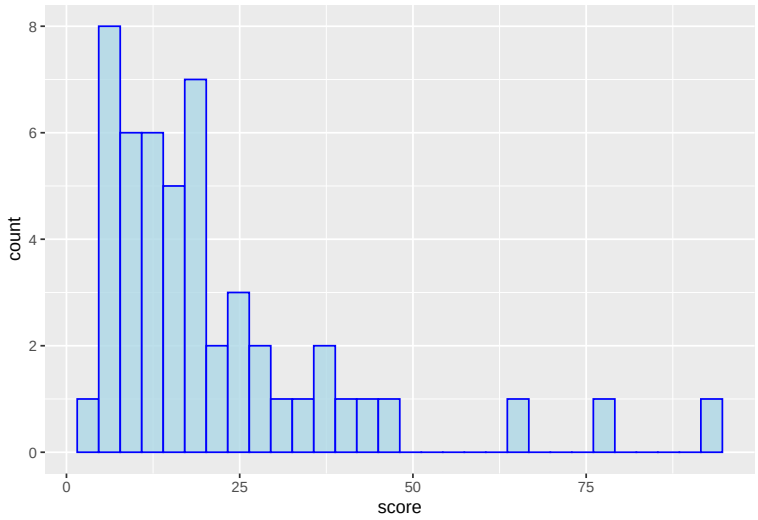
The "Wisconsin Card Sorting Test" is widely used by psychiatrists, neurologists, and neurophycologists with patients who have a brain injury. Patients with any sort of frontal lobe lesion generally do poorly on the test. The data frame WCST contains the test scores from a group of 50 patients from the Virgen del Camino Hospital.

(a) Using the code below, load the WCST data and explore whether there is reason to believe that the score data comes from a non-normal distribution. First, create a histogram (use ggplot!) and describe whether the data look normal. Then, use the function shapiro.test() to explore normality. Be sure to explain what this function does–i.e., what's the null and alternative hypothesis–in your answer. Q18: 2/2

Answer:

2 +1

```
library(PASWR2)
library(ggplot2)
df_b3 <- WCST
ggplot(data = df_b3, aes(x = score)) + geom_histogram(fill = "lightblue", alpha = 0.8,
color ="blue")
```



The data is not look normal.

```
shapiro.test(df_b3$score)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_b3$score
## W = 0.77472, p-value = 2.405e-07
```

Performs the Shapiro-Wilk test of normality.

(b) What assumptions must be made in order to compute a (non-boostrap) 95% confidence interval for the population mean score? Q19: 2/2

Answer: sample mean normally distn 1 0

We have used simple random sampling, or if individuals have been assigned to treatments at random. Ideally our data should be drawn from a normally distributed population.

(c) Compute the confidence interval from (b). Q20: 1/2

Answer: 2 +1

```
library(Rmisc)
CI(df_b3$score, ci=0.95)
```

```
##    upper     mean    lower
## 26.71096 21.48000 16.24904
```

(d) Compute a 95% bootstrap pivot confidence interval for the mean. Q21: 2/2

Answer: 2 +1

```
myData <- df_b3$score
set.seed(1234) # Setting the seed for replication purposes
sample.size <- 50 # Sample size
n.samples <- 50 # Number of bootstrap samples
bootstrap.results <- c() # Creating an empty vector to hold the results
for (i in 1:n.samples)
{
    obs <- sample(1:sample.size, replace=TRUE)
    bootstrap.results[i] <- mean(myData[obs]) # Mean of the bootstrap sample
}
CI(bootstrap.results, ci=0.95)
```

```
##    upper     mean    lower
## 22.15616 21.41120 20.66624
```

1.2.4 Problem B.4

The dataset gives the number of births per month in New York city, from January 1946 to December 1959 . The data are ordered.

(a) Construct another column in the dataset that labels the month and year for each birth per month record. Q22: 2/2

Answer: 2 +1

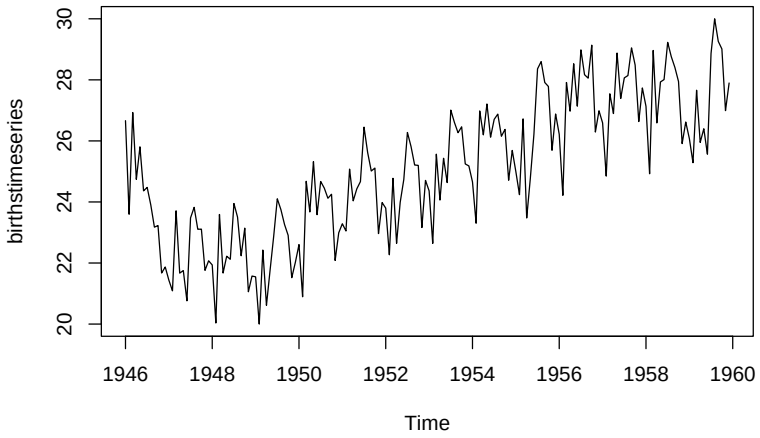
```
library(TTR)
# Loading Data
births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")
birthstimeseries <- ts(births, frequency=12, start=c(1946,1))
birthstimeseries
```

```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 1946 26.663 23.598 26.931 24.740 25.806 24.364 24.477 23.901 23.175 23.227
## 1947 21.439 21.089 23.709 21.669 21.752 20.761 23.479 23.824 23.105 23.110
## 1948 21.937 20.035 23.590 21.672 22.222 22.123 23.950 23.504 22.238 23.142
## 1949 21.548 20.000 22.424 20.615 21.761 22.874 24.104 23.748 23.262 22.907
## 1950 22.604 20.894 24.677 23.673 25.320 23.583 24.671 24.454 24.122 24.252
## 1951 23.287 23.049 25.076 24.037 24.430 24.667 26.451 25.618 25.014 25.110
## 1952 23.798 22.270 24.775 22.646 23.988 24.737 26.276 25.816 25.210 25.199
## 1953 24.364 22.644 25.565 24.062 25.431 24.635 27.009 26.606 26.268 26.462
## 1954 24.657 23.304 26.982 26.199 27.210 26.122 26.706 26.878 26.152 26.379
## 1955 24.990 24.239 26.721 23.475 24.767 26.219 28.361 28.599 27.914 27.784
## 1956 26.217 24.218 27.914 26.975 28.527 27.139 28.982 28.169 28.056 29.136
## 1957 26.589 24.848 27.543 26.896 28.878 27.390 28.065 28.141 29.048 28.484
## 1958 27.132 24.924 28.963 26.589 27.931 28.009 29.229 28.759 28.405 27.945
## 1959 26.076 25.286 27.660 25.951 26.398 25.565 28.865 30.000 29.261 29.012
##         Nov    Dec
## 1946 21.672 21.870
## 1947 21.759 22.073
## 1948 21.059 21.573
## 1949 21.519 22.025
## 1950 22.084 22.991
## 1951 22.964 23.981
## 1952 23.162 24.707
## 1953 25.246 25.180
## 1954 24.712 25.688
## 1955 25.693 26.881
## 1956 26.291 26.987
## 1957 26.634 27.735
## 1958 25.912 26.619
## 1959 26.992 27.897
```

(b) Construct a plot of births per month against the month/year column that you created in part (a). Analyze the plot. Do you notice anything interesting? Q23: 2/2

Answer: 2 +1

```
plot(birthstimeseries)
```



It follows similar pattern each year and most birth given on July.

(c) Suppose that your boss asked you to use the bootstrap to construct a confidence interval for the average number of births per month in New York city over the time period in the dataset. Write a short response to your boss describing why this confidence interval is not valid for this data. Q24: 2/2

Answer:

Bootstrap is used for normal distribution and the births per month in New York city is not normal. If we use the bootstrap, then it result will be vary based on the sample estimator.

bootstrap isn't dependent on normal values or not rld